LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-233798

Data Analysis in the Twenty-First Century

A. Goodman, C. Kamath, V. Kumar

August 17, 2007

Statistical Analysis and Data Mining

DATA ANALYSIS IN THE TWENTY-FIRST CENTURY

The 21st Century is characterized by complex multidisciplinary problems accompanied by massive datasets. "We are drowning in data, but starving for knowledge", as the volumes of many commercial, industrial and scientific datasets have exceeded the terabyte range and are approaching petabytes and beyond. Statistical methodology has long been employed to find useful and usable information in data. More recently, data mining has harnessed the power of computer technology to find useful and usable patterns in such massive datasets.

Although several data mining journals have joined the established statistical journals, no single journal provides an integrated treatment of statistical analysis methodology and data mining technology, particularly when applied to the solution of practical problems. This absence and the needs expressed above motivated the inauguration of John Wiley's new Journal on Statistical Analysis and Data Mining.

The goals of this interdisciplinary journal are to encourage collaborations across disciplines, communication of data mining and statistical techniques to both novices and experts involved in the analysis of data from practical problems, and a principled and productive evaluation of analyses and solutions. The journal specifically encourages submission of works that have statistical rigor in the analysis of data, incorporate the most appropriate algorithms from data mining, and address the needs of applications.

Applying data mining algorithms to practical problems is not sufficient, because we need to ensure that the results have a sound statistical basis, lest any decision based on these results lead to a catastrophe. Even data mining algorithms founded on sound statistical analysis are not sufficient, if they cannot solve a practical problem. Finally, employing a statistical analysis on a practical problem is not sufficient, unless it scales up to massive datasets. Statistical analysis and data mining are actually two sides of the sword that is sorely needed to conquer data overload in practical problems.

Statistical Analysis

Statistical analysis has been used to make sense out of data since King David numbered his people and the Egyptians measured their fields. Quetelet analyzed social data and Galton analyzed biological data in the 1800s, while Karl Pearson, Fisher and Snedecor analyzed agricultural data in the early 1900s.

Then Fisher introduced both mathematics into data analysis and experimental design, Neyman and Egon Pearson introduced statistical inference, and Wald introduced both sequential analysis and decision theory. After World War II, the application of rigorous mathematics to the theory of data analysis has not only flourished, but may have outdone itself by transporting statistical analysis to a position in which it is increasingly difficult to evaluate results against reality. John Tukey's response was exploratory data analysis,

which wisely begins with the data rather than beginning with theory, as had become the custom.

Data has again become dominant in the 21st Century, and now poses both a problem and a solution. The downside of data lies in their volume, while the upside of data lies in an ability to describe reality and to also suggest theory. Similarly, the upside of statistics lies in a capability to make sense out of data, while the downside of statistics lies in the difficulty in scaling its methodology up to handle the overload of data overload.

The ultimate goal of data mining is to transition from exploring data, through exploiting the results, to explaining the data and their results. In order for data mining to achieve that transition successfully, it must pass through the heart of statistics. On the other hand, statistical analysis must be rigorously scaled up to the level of data volume.

Many 21st Century opportunities and challenges for statistical analysis lie in the effective management and compression of massive datasets, motivation and justification of data mining algorithms, support of the transition from data exploration to data and result explanation, and evaluation of data mining results against reality. In addition, statistical analysis may well be useful in creating value from data mining results by yielding new insights, motivating decisions, and justifying actions.

Data Mining Algorithms

Algorithms for analyzing data have been studied by statisticians and used in a variety of disciplines dating back many centuries, but new algorithms need to be designed to address the limitations of the existing techniques in handling the challenges posed by the new types of data that are now being collected. Recent advances in information technologies have enabled collection of vast amounts of data in commerce and a variety of scientific disciplines. Many of these data sets have one or more of the following characteristics: high-dimensional, heterogeneous, distributed, streaming, spatio-temporal nature, etc. Often traditional techniques cannot be applied to these data sets either due to their massive size or due to their non-traditional nature.

The emerging field of data mining grew out of the limitations of the existing data analysis techniques for addressing these new types of data. In a short period of a little over 10 years, the data mining community has grown rapidly and continues to produce a large number of algorithms that try to address the above limitations. While many of the algorithms being developed by this community build upon a long history of work in other disciplines (e.g. those for classification, clustering, outlier detection), some algorithms (e.g., association pattern analysis) are unique to the field.

There are a number of challenges in developing an effective algorithm. Often the data collected is low-level and it is necessary to extract higher level features before a mining algorithm can be applied. Often times, association pattern analysis algorithms produce a

very large number of patterns and determine which of these patterns is useful can be very challenging.

Addressing these challenges requires close interaction between data mining algorithm designers and application experts since deep knowledge of the domain the key to identifying and validating useful high-level features as well as determining patterns that are meaningful. Data mining algorithms by their very nature are data driven. The hypotheses generated by these algorithms must be validated by sound statistical methodology for them to be useful in practice.

Applications

The main driving force behind the development of statistical techniques and data mining algorithms is the data analysis problem which needs to be solved. The domains which give rise to such problems are as varied as the solution approaches. For example, an astronomer may be interested in automatically identifying the shape of the galaxies in an astronomical survey to understand if the existing models of the universe are correct. Or, a computational biologist may be generating petabytes of data on a massively parallel machine, simulating the working of the human brain - data which must be analyzed to acquire this insight. The problems being addressed can have a broad influence, affecting the day-to-day lives of many, such as the correct identification of a fraudulent credit-card transaction or the return of relevant results in response to a search engine query. These are but a few examples where analysis of data is becoming an integral part of our lives. And the list continues to grow as computer simulations are being used to understand complex phenomena ranging from turbulence and the properties of materials to the spread of disease and the structure of proteins.

There are several factors which make the analyses of these data challenging, and therefore interesting and worthy of pursuit. In some problems, the domain expert may be looking for something which is ill-defined. In other problems, the data may be been collected with one goal in mind and is now being analyzed for a different purpose. It could have missing values, be noisy with a low signal-to-noise ratio, be high-dimensional, or be unstructured. The domain expert may want the data to drive the analysis so that the results are not influenced by what they expect to see in the data. Or, they may need assurance that the results are a true reflection of the data rather than an artifact of the algorithms or parameters. In some cases, they may want certain domain knowledge to be incorporated into the analysis such as the cost of a false negative.

These challenges may appear insurmountable and the problem domains so varied that we may think that solutions are beyond our reach. However, the variety of problems may be a blessing in disguise. It is likely that the problem of interest to one application specialist may have already been solved in a different domain, if not completely, at least partially so one can get some ideas on how to proceed. It is these aspects that we hope to bring to our readers in the area of applications - solutions to challenging problems and the opportunity to learn from other application areas similar to one's own.

Concluding Remarks

The Statistical Analysis and Data Mining journal focuses on three areas:

- statistical analysis,
- data mining algorithms, and
- practical applications

indicating the importance of each area separately and the interaction of all the three areas in ensuring a successful data analysis effort.

We welcome papers in each of these areas. However, our goal is to motivate authors to see the benefits of expanding their focus to include the other areas. For example, we would like the algorithm developers and statisticians to realize the value of incorporating each other's techniques in their own work. We also want to encourage them to apply their techniques to the solution of real problems, in the process enhancing their techniques to handle the idiosyncrasies of real-world data. Similarly, we would like those analyzing data from real applications to share their experiences so others in related areas may benefit, or an algorithm developer or statistician may propose a solution.

As this first issue of Statistical Analysis and Data Mining goes online and into print, we hope you, the reader, will interact with us and contribute to our endeavor. We welcome your comments on the content and format of the journal, as well as your submissions. We sincerely hope that this, and future issues, will help provide solutions to your data analysis problems.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Arnold Goodman,
Chandrika Kamath, and
Vipin Kumar
Editors-in-Chief